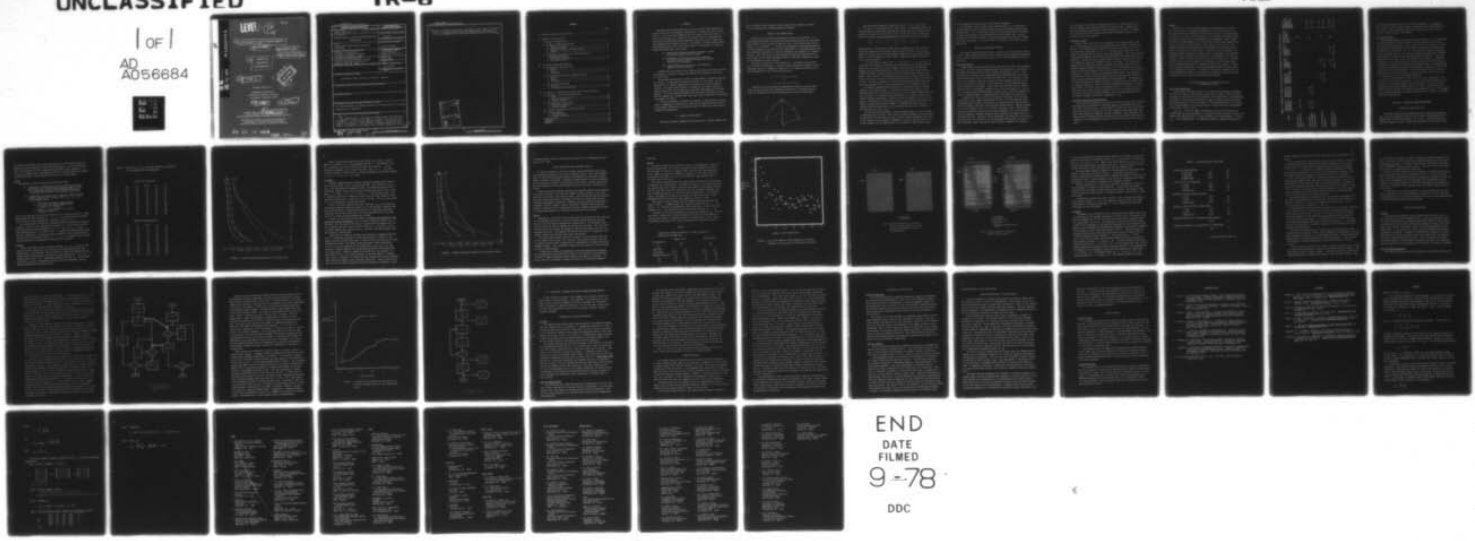
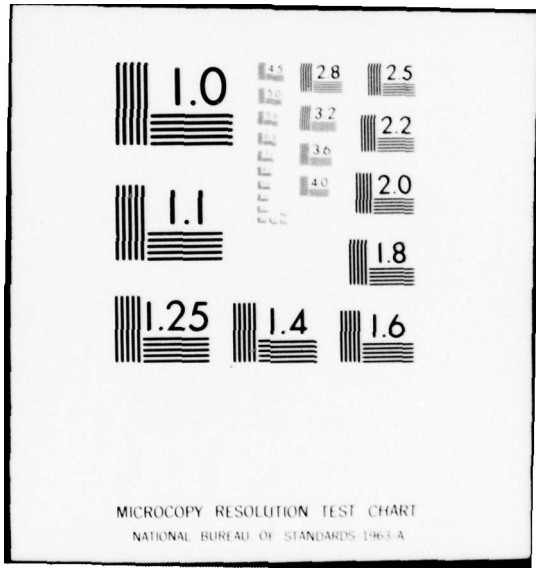
AD-A056 684 UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES DEPT O--ETC F/G 5/10
IMPLIED ORDERS AS A BASIS FOR TAILORED TESTING.(U)
APR 78 N CLIFF, R CUDECK, D MCCORMICK N00014-75-C-0684

UNCLASSIFIED TR-6 NL

1 OF 1
AD A056684



END
DATE FILMED
9-78
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

LEVEL II

12

(6) IMPLIED ORDERS AS A BASIS FOR TAILORED TESTING.

(9) FINAL REPORT,

(16) RR042Ø4

(17) RR042Ø4Ø1

(10) Norman Cliff,
Robert Cudeck
Douglas McCormick

(14) TR-6

Technical Report No. 6

Department of Psychology
University of Southern California
Los Angeles, California 90007

(11) Apr[redacted]78

(12) 42p.

(15) N00014-75-C-0684
Prepared under contract No. N00014-75-C-0684
NR No. 150-373, with the Personnel and
Training Research Programs, Psychological Sciences Division

400 762

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report No. 6	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* IMPLIED ORDERS AS A BASIS FOR TAILORED TESTING FINAL REPORT		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Norman Cliff, Robert Cudeck and Douglas McCormick		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0684
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology University of Southern California Los Angeles, California 90007		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61153N, RR042-04, RR042-04-01, NR 150-373
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Program Office of Naval Research (Code 458) Arlington, Virginia 22217		12. REPORT DATE April, 1978 13. NUMBER OF PAGES 38
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

(U) Tailored Testing; (U) Adaptive Testing; (U) Testing Theory; (U) Internal Consistency; (U) Simulation

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The research from a 4 year project of implied orders tailored testing is summarized. Included in the report is a review of all work previously carried out, as well as recently completed research. Three topics were studied in this project: the production of computer programs for tailored testing, contributions to test theory from an ordinal perspective, and a

over

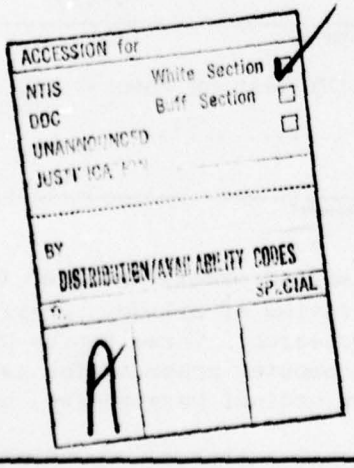
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-LF-014-6601

-A

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

pattern for sound research which is applicable for all types of tailored testing. An overall evaluation of the progress made in each area is given and recommendations for additional research projects are included.



CONTENTS

I. OVERVIEW

This paper is the final report of a four-year project which examined tailored testing from a non-parametric perspective. Its most distinctive characteristic is the use of the Guttman scale as an ideal of the test setting. The goal of testing from such a viewpoint is a mutual ordering of examinees and items along a hypothetical ability dimension, making use of only simple counting procedures to accomplish the task. This method is called tailored testing using implied orders, and in the current project the development of these procedures progressed along three fronts.

(1) Producing and evaluating computer programs to accomplish implied orders tailored testing.

(2) Development of principles of ordinal measurement which will provide a theoretical basis for tailored testing and its evaluation.

(3) Suggest methods for carrying out and evaluating research in tailored testing.

Although this third aspect of our work has more often been implicit rather than explicit, we believe it may be an important contribution for the field in its own right.

These three general concerns, namely the development of computer algorithms for testing, the advancement of ordinal techniques in test theory, and the delineation of a research methodology for tailored testing have guided our research. The report which follows is a summary of our current thinking in each of these areas, and includes many conclusions and suggestions for future research. It is our overall impression that the method developed so far has much to recommend it, although after only three years a great deal of additional work remains to be done.

It will be convenient to present this material in several major sections. First, a comprehensive review will be given of the research carried out prior to this paper, after which will follow a report of recent data, and finally an evaluation and prospectus.

II. Summary of Past Research

This section briefly summarizes the progress made on Implied Orders test-

ing up to recent times, includes a short review of our research in several areas, and sketches the rationale behind the methods.

Review of the TAILOR System

The test tailoring system developed here has been called TAILOR. It is a simultaneous procedure for acquiring information about person and item orders, which unlike all other methods of tailoring tests, treats persons and items as elements of a joint order. It is based on an approach of simultaneously gathering item information concurrent with the administration of tailored testing, and thus eliminates the need for a block of examinees to be pretested on a complete item pool. This elimination of pretesting may one day bring tailored testing within reach of test givers who cannot afford to administer thousands of tests as a preliminary to actual tailoring.

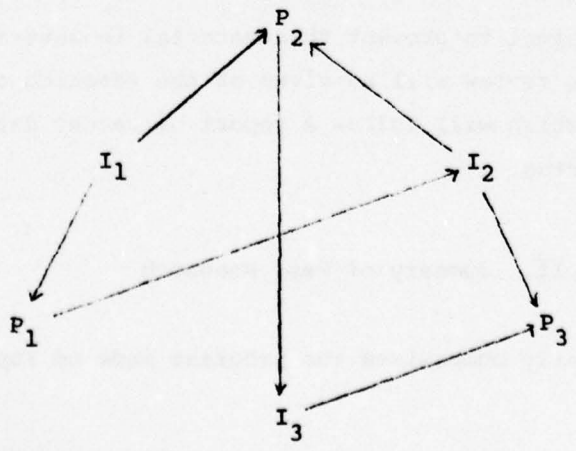
The basic principle is a rather simple one. Dichotomous items provide only two kinds of information. Either an item is missed and therefore, in the terminology used here, dominates a person or it is answered correctly and the person *dominates the item*. Depicted pictorially,

$$I \longrightarrow P$$

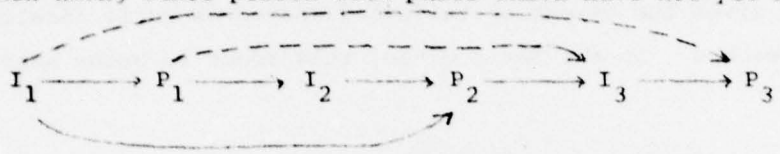
or

$$P \longrightarrow I$$

Generally we have discussed the mathematics of TAILOR in matrix terminology, but as has been stated elsewhere (Cliff, Note 2) these matrices are merely the formalization of directed graphs such as the one shown below.



Each solid arrow represents an observed correct ($P \longrightarrow I$) or incorrect ($I \not\longrightarrow P$) answer on a hypothetical test. When placed in a reasonable order such as shown below, the above graph demonstrates how the observed relations between persons and items can first be used to give an order among items or among persons and then among other person-item pairs which have not yet been observed.



When presented in this manner it should also be apparent that if a consistent pattern of dominances exist between nearly adjacent elements, then this also implies the nature of the relations between elements which are located farther apart. It is this ability to imply distant, unobserved relations that makes test tailoring possible.

If one long arrow is added from I_1 on the left to P_3 on the right, as indicated by the dashes, the overall order relationships are not much enhanced. Every test user knows that it is of doubtful utility to pair the brightest examinee with the easiest questions. But the situation for closely matched persons and items is likely to be less clear cut and therefore much more informative. When a welter of relationships indicates that no simple order exists then a statistical evaluation of the weight of evidence may serve to clarify the picture. Just such an approach is taken by TAILOR, and the details have been described elsewhere (Cliff, Cudeck and McCormick, Note 4; McCormick, Note 6)

It should be emphasized that although the persons and items are already ordered in the directed graph, in a matrix they need not be ordered for the implication process to take place and in fact are not formally ordered until the end of testing. So, although formal pretesting is not necessary, TAILOR does require responses from people to order its items and thereby to also order its people. The operation of this process occurs in two possible ways which correspond to two separate versions of TAILOR: a FORTRAN group testing version (Cudeck, Cliff and Kehoe, 1977) and an individual testing version written in APL (McCormick and Cliff, 1977).

In the group testing approach, several examinees begin at once to attempt a random selection of items. From a very few initial responses the order begins to take shape, and tailoring can occur to varying degrees depending on the size of the item pool and the number of examinees. No one in the group version is given a complete set of items and there is no problem of item security caused

by early examinees discussing the test with later examinees.

The individual testing version gives complete tests one at a time until enough information exists to begin eliminating the most extreme items from the tailored tests. In the tests given to date with this second method, the number of persons tested before tailoring begins has been invariably two. As more examinations are given the extent of tailoring increases until finally an asymptotic value is reached. In the tests given, this seems to occur near fifteen examinations.

Review of Previous Evaluations

Four types of evaluations have been carried out on the Implied Orders model to date. Rather complete reviews of this research have been presented previously (Cliff, Cudeck, and McCormick, Note 3; Note 4), but a brief summary will also be given here.

Monte Carlo Results

A Monte Carlo study with errorless data was the first step in the evaluation of TAILOR. The group testing program was used for this project, the details of which have been given (Cudeck, Cliff, Reynolds and McCormick, Note 5). With errorless data this Monte Carlo demonstrated that TAILOR was able to perfectly order the persons and items, and needed only about 50% of the responses. Departures from a complete order only occurred in those instances where the "true scores" assigned to two persons (or two items) were so close together that no item true score fell in between. Knuth (1973) has indicated that the minimum number of bits of information required to sort a set of n elements is $\log_2 n!$. It is interesting that TAILOR performed a similar task to sorting elements with nearly the minimum number of responses given by the theoretical expectation.

A second type of Monte Carlo was carried out which used a more realistic kind of data (Cliff, Cudeck and McCormick, Note 4). With the Birnbaum (1968) 4-parameter logistic model, various sizes of response matrices were generated which were based on specified item and person characteristics. Samples included 10, 25 or 40 "persons" and 15 or 25 "items", each had at least 5 replications and most were based on several sets of latent parameters. As in the errorless case, the number of responses required varied inversely with sample size, but averaged around 50%. With this kind of data model, the correlation of true score with test score is a validity in the classical test theory sense. By far the most significant effect on tailored tests validity is complete test validity

(the correlation of ability with complete test score). Tailored validity was close to complete test validity although a few points lower. However there was an increasingly large effect in tailored validity as compared to complete test validity, such that as complete test validity decreased, tailored test validity fell off more rapidly.

Simulation Studies

A file of data consisting of the responses to 122 Stanford-Binet items by 622 children was used in the next evaluation (Cliff, Cudeck, and McCormick, Note 4). The children ranged in age from 2 years to just under 15 years, and all together had an average IQ of 117.3 . The sample was divided into 3 age groups and simulations took place on the groups separately. The method used here is especially important for tailored testing research and deserves careful presentation. For any age group a random sample of subjects was selected, and for this sample the items were randomly split into two halves. The correlation between the two data sets was used to compute a reliability which was called Complete-Complete parallel forms reliability. One of these matrices was then input to the testing program, which used the data as the basis for a simulation, operating in a manner similar to the previous Monte Carlos. The correlation between the score on the tailored matrix with the score from the other half of data produced a Complete-Tailored parallel forms reliability. The comparison of Complete-Complete with Complete-Tailored reliabilities was the primary means of comparison used here, and it is the most straightforward method of evaluation possible (see Cliff, Cudeck and McCormick, Note 3). As with the fallible Monte Carlo data, about half the responses were used, the Complete-Tailored reliability was close to Complete-Complete, and it was quite dependent upon the Complete-Complete reliability, irrespective of age.

Individual Testing with Human Subjects

The above Monte Carlos were all carried out with the group testing program. McCormick (Note 6) developed an individual approach, and produced the first data based on human subjects. Using anagrams, he found that the first few subjects took nearly the entire test, but that subsequent ones needed fewer and fewer items reaching an asymptote of about 9 questions. More important, the parallel forms reliability for two tailored tests was actually higher than the reliability for an independent sample who took the same items as two complete tests.

Summary

In Table 1 is a much distilled summary of the research done to date. The various methods of evaluating performance are listed across the top of the table and the four different designs are given as rows. Although slight variations exist between outcome measures, the consistent philosophy has been to compare tailored testing performance with an independent sample of data. The various outcome measures in the last column are quite high, ranging from .93 to 1.07, while the percentage of items used went from 44 to 55. The summary findings are uniformly positive and suggest that the Implied Orders model may be useful for several kinds of tailored testing requirements. It should be indicated, however that these results are averages, that sometimes the method can be expected to perform better, but sometimes it may actually do worse. Of most concern to us is the fact that the accuracy of a tailored testing session is a direct function of the quality of the original data. Since this method begins with no information about either subjects or items, it is especially susceptible to producing tailored results that are invalid as a consequence of poor data. Our tentative evaluations regarding the testing algorithms are linked to an evaluation of the quality of data, such that the better the data, the better the performance has been.

Consistency Evaluation and the Identification of Unidimensional Orders

Consistency Measurement

In the context of Implied orders, test items provide dominance information. The number of times an item dominates another item is readily found by pre-multiplying the rights matrix by the transpose of the wrongs matrix. Correspondingly, the number of times one person dominates another is found by carrying out that same multiplication in the reverse order. If the dominance matrix is consistent, it is highly asymmetric, and several measures of such consistency were proposed (Cliff, Note 2; Cliff, 1977).

One, c_{t2}, counts the number of dominance relations that are above the diagonal when the item=item matrix is in difficulty order, divides it by the total number, and transforms this to a more reasonable scale (See Appendix). A second, c_{t3}, is more sophisticated. It compares the number of dominance relations to

TABLE 1 : Measures of Complete and Tailored test performance from four previous studies.

Study	Average Correlation of Tailored and True Scores	Average Correlation of Complete and True Scores	Average Correlation of Complete and Tailored Scores	Average Reliability of Complete Scores	Average Reliability of Tailored Scores	Proportion of the Item Pool in Tailored Tests	Ratio of Measures of Tailored to Complete Test Performance
Monte Carlos Using Errorless Data	Tau = .96					.48	
Monte Carlos With 4-Parameter Logistic Model	r = .89 Tau = .76	r = .93 Tau = .81				.55	$\frac{r}{r} = .94$ $\frac{r}{r} = .96$
Stanford-Binet IQ Simulation			r = .81 Tau = .66	r = .84 Tau = .71		.51	$\frac{r}{r} = .93$ $\frac{r}{r} = .96$
Anagram Data with Human Subjects				r = .78 $Tau_B = .61$	r = .83 $Tau_B = .65$	.44	$\frac{r}{r} = 1.06$ $\frac{r}{r} = 1.07$

the number that would occur if the items were independent. The symmetric way in which the theory treats persons and items allows parallel indices for consistency of person dominance relations, although the two types are not directly related except at the extremes.

Multidimensionality

One promising answer to the need for pretesting may lie in the derivation of ordinal measures of item consistency, used in conjunction with on-line tailored testing in real time. What we have had in mind is similar to a factor analysis of dichotomous items, the need for which has not escaped other researchers (eg. Bock, Note 1, p. 47). The first step toward this goal was taken by Reynolds (Note 7), who outlined a method of constructing a unidimensional sample of items from a larger heterogeneous population. The basic proposition in this work is that the Guttman scale can serve as a model for item selection if one has a means of assessing the degree to which internal consistency is improved by the addition of candidate items. Reynolds applied a measure of consistency which has been advanced by Cliff (Note 2, 1977), and his results were quite encouraging. Of course his work does not include an application to the testing setting in exactly the form called for here, but it is important as a first step toward that goal.

However, the usefulness of these ideas of internal consistency is not limited to construction of item pools. In fact the primary advantage of such an index is as an omnibus measure of overall quality of a dataset. Since adaptive testing research demonstrates such a direct relation between the quality of test items and the accuracy of the examinee's scores, it is apparent that consistency measures can provide information which is vital to the understanding of testing data.

SECTION III. PREVIOUSLY UNREPORTED RESEARCH

Consistency Index Monte Carlo

The indices which have been proposed actually have quite broad application, and in the testing field may be used in cases of both complete and incomplete data. Two of these indices, c_{t2} and c_{t3}, seem especially promising for adaptive testing. Because the measures are of such recent origin

and have not had extensive use outside this project, a brief study of their behavior under known conditions of data seemed appropriate. In the following Monte Carlo the performance of some standard testing statistics were compared to c_{t2} and c_{t3} with complete data. The Appendix provides a thorough discussion of computational procedures for both measures and also gives a worked example.

Method

The following 2-phase procedure was used in this study:

(1) A theoretical response matrix of 300 persons and 200 items, with ability and difficulty distributional parameters fixed, chance values set to zero and mean discrimination equal to μ_a was generated according to the Birnbaum (1968) model.

(2) A sample matrix of p persons and i items was randomly selected from the larger matrix. Values of KR20, c_{t2}, c_{t3}, v_r and v_t were computed, where

v_r = the Pearson correlation between ability true scores and obtained number correct in sample (i.e., validity)

v_t = same as v_r, only Kendall's Tau is used instead of Pearson r.

Step (2) was executed 10 times to get mean scores on the above statistics. These means became the values for KR20 μ_a, c_{t2} μ_a, etc. which are later reported. Steps (1) and (2) were performed for a range of μ_a from 0.2 to 2.0.

The fixed ability and difficulty parameters used here described two theoretical tests. The first is an optimal test situation, i.e., with ability and item difficulty having the same parameters, $\mu_b = \mu_\theta = 0$ and $\sigma_b = \sigma_\theta = 1$, both from normal populations. As usual, b and θ describe item difficulty and person ability, respectively. The second situation represents a broad range difficulty test which is quite difficult with $\mu_b = 1$, $\sigma_b = 2$. Henceforth these theoretical populations will be designated as "Optimal" and "Difficult".

Results

Table 2 and Figure 1 shows the values for KR20, c_{t2}, c_{t3}, v_r and v_t (the metric and ordinal validities) as a function of item discrimination for the Optimal tests. As can be seen, the two indices of validity are moderate for the lowest values of discrimination, and begin to reach asymptote for $\mu_a \geq .80$. As expected KR20 is about v_r^2, asymptoting slightly later than v_r. On the other hand, both c_{t2} and c_{t3} increase monotonically across discrimination, and c_{t3} appears nearly linear as a function of μ_a.

TABLE 2 : Values of c_{t2} , c_{t3} , KR 20 and validity as a function of discrimination for two populations

Optimal Test Population

μ_a	c_{t2}	c_{t3}	KR 20	v_r	v_t
.2	-.54	.03	.34	.59	.43
.4	-.22	.11	.62	.79	.62
.6	.04	.24	.78	.88	.72
.8	.16	.35	.85	.92	.79
1.0	.32	.47	.89	.93	.81
1.2	.44	.53	.90	.94	.82
1.4	.54	.60	.91	.94	.84
1.6	.74	.68	.91	.95	.84
1.8	.75	.77	.93	.96	.88
2.0	.76	.79	.94	.96	.88

Difficult Test Population

μ_a	c_{t2}	c_{t3}	KR 20	v_r	v_t
.2	-.34	.04	.39	.63	.46
.4	.19	.10	.53	.76	.59
.6	.49	.24	.71	.86	.71
.8	.67	.36	.77	.87	.72
1.0	.74	.46	.81	.90	.76
1.2	.82	.50	.79	.90	.78
1.4	.83	.66	.87	.93	.82
1.6	.89	.69	.85	.93	.82
1.8	.92	.72	.84	.92	.80
2.0	.94	.74	.83	.93	.83

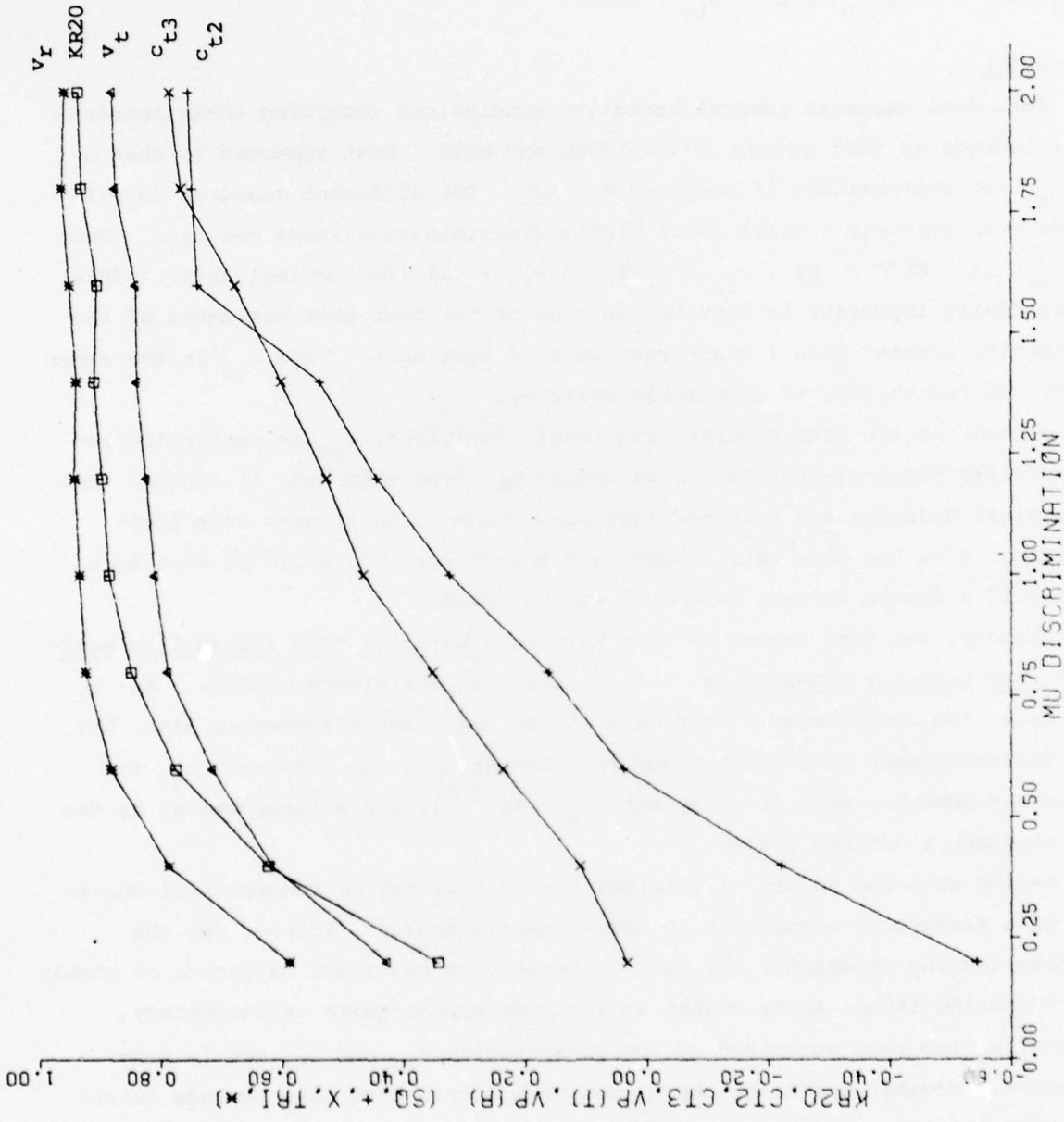


FIGURE 1: Internal consistency measures for the Optimal Tests.

Figure 2 contains the results for the Difficult test. Again, validity begins at a moderate .4 or .6 and asymptotes around .8 . KR20 behaves as before, although it more closely follows the rank order validity. c_{t3} also displays an asymptote, but it occurs much later, near $\mu_a = 1.4$. c_{t2} shows a regular, monotonic increase which is actually greater than the metric validity at $\mu_a = 2.0$. For $\mu_a > 1.6$, $c_{t3} >$ KR20 .

Discussion

This data suggests several tentative conclusions regarding these consistency indices as they relate to validity and KR20. Most apparent is that c_{t2} and c_{t3} are conservative if compared to KR20. The different measures do not become even remotely similar until highly discriminating items are used. Thus when $\mu_a = .4$, KR20 = .62 , $c_{t2} = -.22$ and $c_{t3} = .11$ (for Optimal data). This is especially important to consider in view of the fact that estimates of discrimination greater than 1.0 are rare in real test data. Thus c_{t3} in the range .25 to .35 corresponds to acceptable validity.

Second, as the test departs from ideal conditions, c_{t2} in particular behaves fairly regularly across the values of μ_a. From this data it appears that the ordinal measures are somewhat less susceptible to departures from ideal conditions than the more metric KR20, and the discrepancy would be even more drastic if a chance success parameter were allowed.

Thirdly, the data remind us that KR20 is a parallel form reliability estimated from internal consistency, not an internal consistency measure. As μ_a increases, the consistency of the data in the usual sense increases also, but KR20 increases much more quickly and then levels off. In contrast, c_{t2} and c_{t3} behave somewhat more linearly with μ_a , and only reach large values as the data approach a Guttman scale.

Seeing such low values of internal consistency may be somewhat disconcerting to a researcher accustomed to using other measures. Indeed, for the adaptive testing researcher who goes to lengths to construct batteries of highly discriminating items, these modest results may appear quite extraordinary. In fact, we too were surprised at how conservative c_{t2} anc c_{t3} are in many instances. Roughly speaking, the c_t measures behave more like average inter-item correlations. However it should be indicated that the evaluation of a score matrix by the c_t's conforms precisely to the ideal model of a Guttman scale, and it is difficult to imagine any more appropriate basis for evaluation under nonmetric assumptions than a Guttman simple order. These statistics provide for the first time a rationale of internal consistency formulated in terms

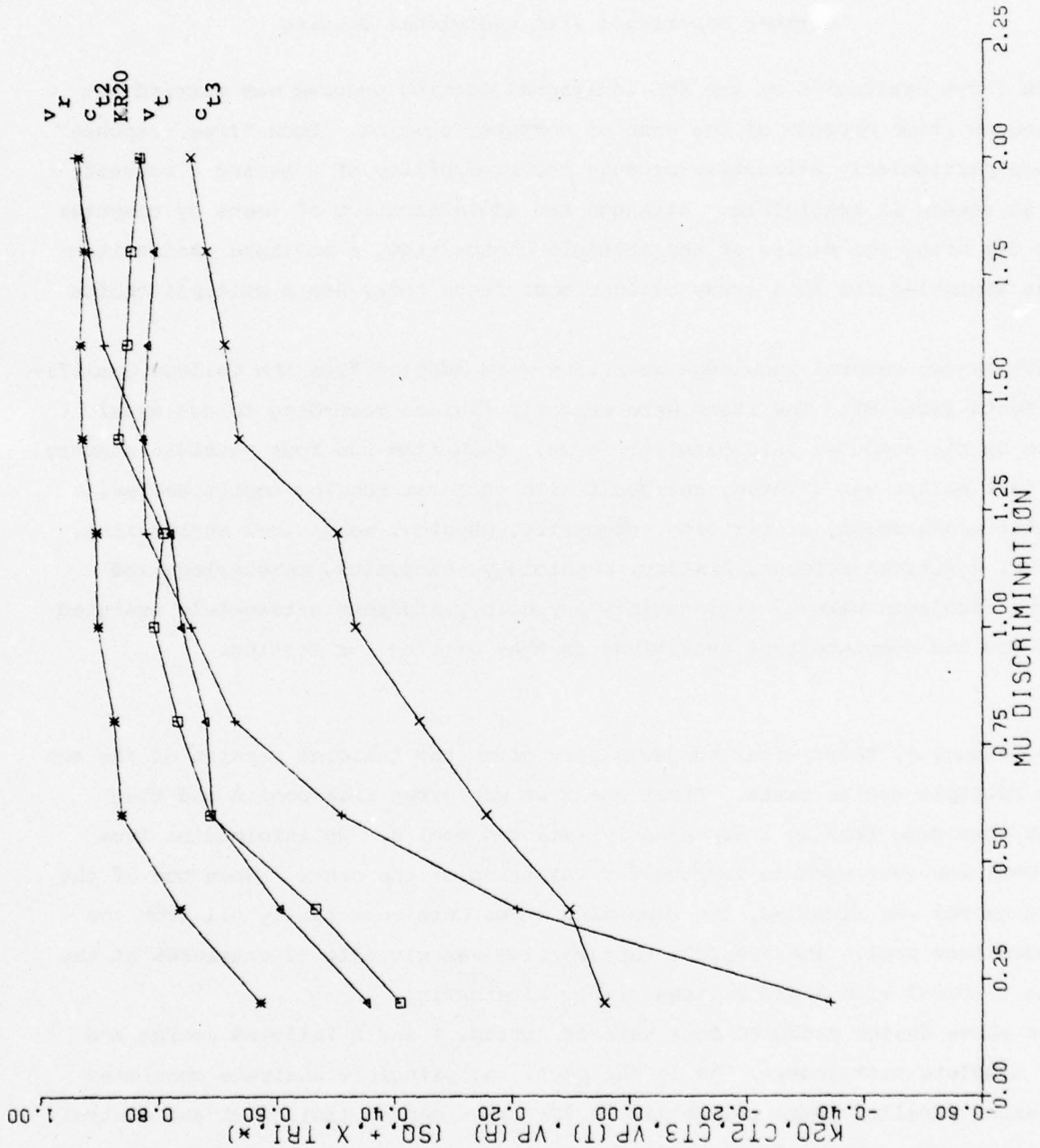


FIGURE 2: Internal consistency measures for the Difficult Tests.

of ordering theory, which in our view is the correct way to conceptualize consistency of test items.

Further Experience with Individual Testing

The first evaluation of the APL individual testing program was carried out with anagram items because of the ease of computer scoring. Such "free response" items are particularly attractive because the probability of guessing a correct answer at random is negligible. Although the administration of tests by computer may one day bring the demise of the multiple choice item, a multiple choice item pool was assembled for this study because most tests today use a multiple choice format.

Seventy-two general knowledge questions were adapted from the College Qualification Tests (Note 8). The items were randomly divided according to our usual practice by the computer into parallel forms. Each item had four possible answers. The subject matter was diverse, and dealt with such far ranging topics as law, scientific measurement, electricity, chemistry, physics, mechanical engineering, geography, political science, history, physiology, economics, meteorology and agronomy. Subjects were 67 introductory psychology students alternately assigned to tailored and complete test conditions as they arrived for testing.

Method

Individually, thirty-four subjects were given the tailored version of the two 36 item multiple choice tests. First one item was given from pool A and then the next item came from an independently tailored pool B. No information from either pool was ever used to improve the tailoring of the other. When one of the tailored halves was finished, the remaining items were necessarily all from the unfinished item pool. The complete test version was given to 33 examinees at the computer terminal with A and B items simply alternating.

The above design produced four sets of scores, A and B tailored scores and A and B complete test scores. As in the past, the principle analysis consisted of comparing parallel forms reliabilities for experimental (tailored) and control (complete) groups. Also of principle interest was the percent of items necessary to obtain tailored scores.

In addition to reliability comparisons, score distributions were examined for irregularities and the sequential nature of the individual testing program utilized to compute regression analyses based on the extent of tailoring. c_{t2} and c_{t3} were also calculated to compare the quality of tailored and complete

test data.

Results

The testing sessions, as revealed in Figure 3, proceeded much as they did for the anagram items. The first two tailored subjects answered all items, and the amount of tailoring increased rapidly over the next dozen administrations, seeming to approach an asymptote of around 11 or 12 items (out of 36) by the end of the study. The overall percentage of items presented was 41 for form A and 42 for form B.

The actual response matrices for complete and tailored tests are presented in Figures 4 and 5 respectively. The items and persons in the tailored matrices are labelled such that a person's identification number in the sequence of testing is at the left of his row. Entries at the side and bottom represent overall net dominance scores for items and persons, respectively. A net dominance score is the number of persons and items that the specified person or item dominates minus the total that in turn dominate the person or item.

The Pearson r and t_B reliabilities for complete and tailored scores are given in Table 3. Disappointingly, the tailored reliabilities were $r = .26$ and $t_b = .19$, whereas the corresponding coefficients were .68 and .57 for the complete tests. The reliabilities are significantly different at $\alpha = .05$.

Tailored test scores corresponding to person-item relations, i.e., the number right or implied right, were calculated also so that the distribution of tailored scores could be compared with that for complete test scores. Means and standard deviations of the score distributions are presented in Table 3.

TABLE 3

Score Means, Standard Deviations, c_{t2} and c_{t3} values for Complete and Tailored Data.

	Complete Data		Tailored Data	
AB reliability				
r	.68		.26	
tau	.57		.19	
Item Pools	A	B	A	B
Score Means	21.39	19.97	24.26	20.32
Standard deviations	5.73	4.54	7.24	6.86
c_{t2}	.04	.10	.32	.32
c_{t3}	.16	.09	.26	.19

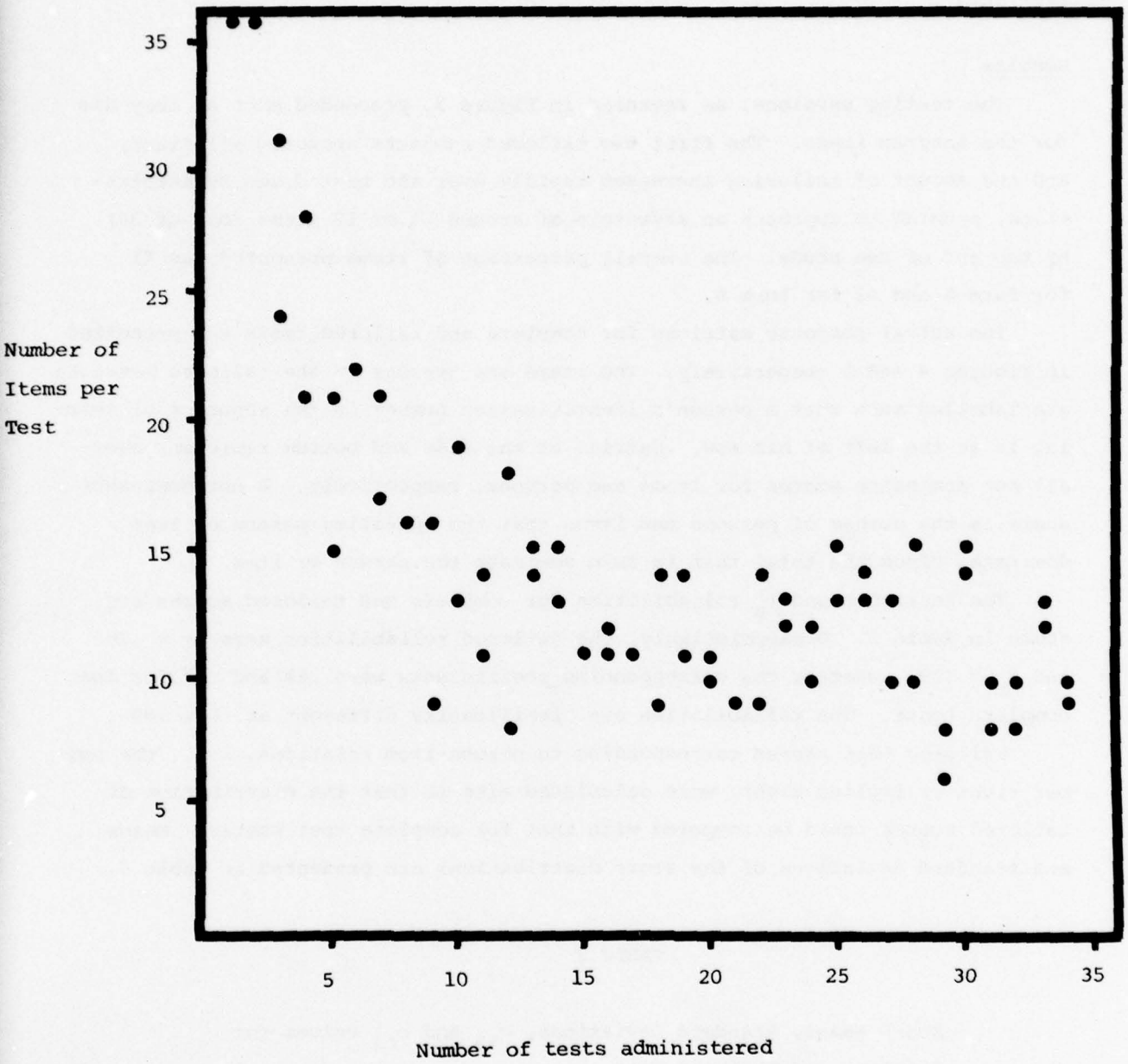
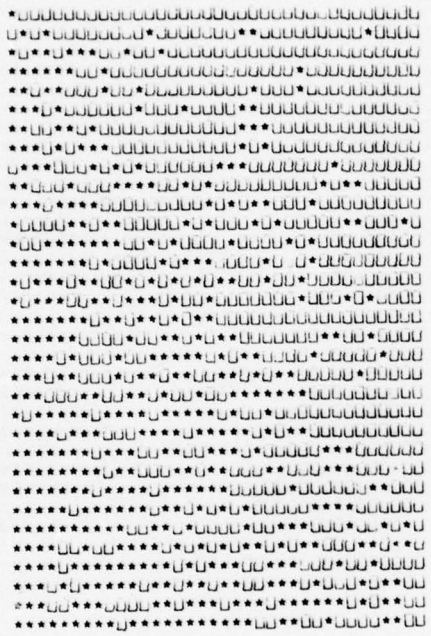


FIGURE 3: Plot of the number of items presented in each test according to its position in the sequence of testing.

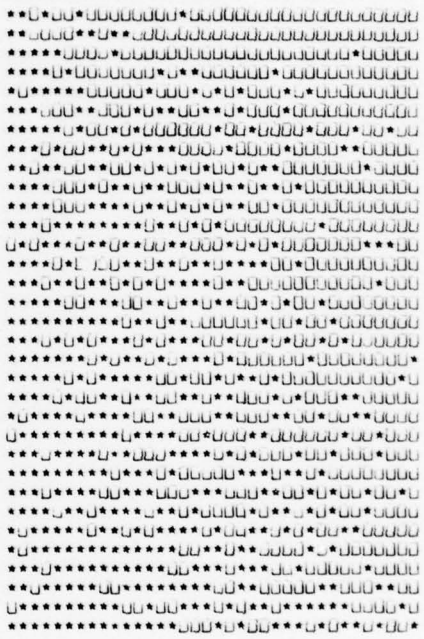
A ITEMS

B ITEMS

PERSONS



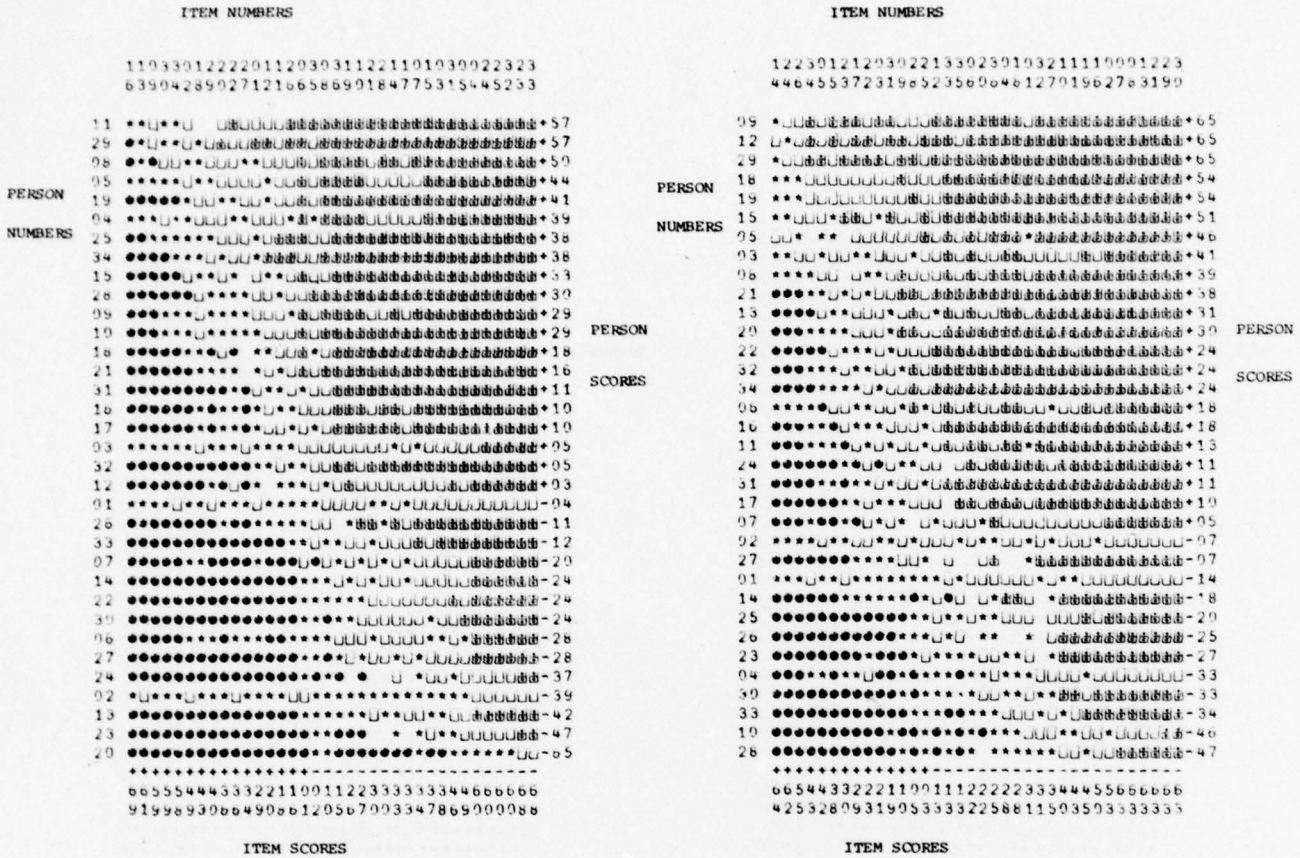
PERSONS



□ Correct answers

* Incorrect answers

FIGURE 4: Complete Test Response Matrices. Rows and columns in the figures are permuted, independently, so they do not correspond to the same persons within figures or the same items between figures.



- Correct Answers
- ★ Incorrect Answers
- ⊞ Implied Correct Answers
- ● Implied Errors

Blanks are implications which were later revoked.

FIGURE 5: Tailored Test reponse matrices. Individuals can be identified by the numbers at the left.

A t-test was performed for tailored scores computed in this manner, and also for complete test scores for item pools A and B. The α levels of the two comparisons were $\alpha_A = .08$ and $\alpha_B = .80$. Table 3 also gives values for c_{t2} anc c_{t3}, for tailored and complete tests, and item pools A and B. As can be seen the tailored test values are in each case higher than those for complete tests.

The relationship between the number of items presented in each test and the number of tests given was shown in Figure 3. Correlations obtained for a variety of measures and serial positions of examinee and number of items taken are presented in order to clarify changes in test scores as a function of the extent of tailoring. The resulting correlations are presented in Table 4.

Absolute differences in z scores of a particular individual on items A and B were examined for trends in score reliability. Correlations with neither serial position nor number of items taken were significant. The absolute values of the z scores were also looked at for possible trends in score variance. None were found.

A significant correlation between z scores in the A item pool and number of items taken was found ($r = -.35$). The combined correlation for A and B items also produced a significant correlation ($\alpha < .05$), although the B item pool correlation was not significant.

Discussion

The complete test reliability of the items used here is somewhat lower than the reliability of the anagram items used in the first study. The Pearson r and τ_B for the anagrams are .78 and .61, while the values for the information items are, respectively .68 and .57. Such decreases in reliability have been shown, in Monte Carlo evaluations of the group testing program to have disproportionate effects on the reliability of tailored tests compared to the decrease in complete test reliability (Cliff, Cudeck and McCormick, Note 4). Regression equations calculated from the Monte Carlo data can be used to predict tailored test reliability from complete test reliability. Given complete test $r = .78$ and $\tau_b = .61$ for the first study, the regression based on the group testing Monte Carlo predicts tailored $r = .66$ and $\tau_b = .49$. The values obtained from individual live testing are, $r = .83$ and $\tau_b = .61$. For the anagram data, then, the obtained results are considerably above the predicted values. If the computations are repeated for the information items which have complete test reliabilities, $r = .68$ and $\tau_b = .57$, the predicted values are, $r = .50$ and $\tau_b = .49$. The obtained values are, $r = .26$ and $\tau_b = .19$. The obtained values for the less

TABLE 4: Correlations from z score data

z score and serial position		n
A item pool	-.161	34
B item pool	-.045	34
Combined data	-.103	68
z score and no. of items taken		
A items	-.349 *	34
B items	-.180	34
Combined data	-.265 *	68
Absolute z scores and serial position		
A items	.014	34
B items	-.035	34
Combined data	-.010	68
Absolute z scores and no. of items taken		
A items	.002	34
B items	-.052	34
Combined data	-.024	68
Absolute difference in z scores and serial position		
	-.124	34
Absolute difference in z scores and no. of items taken		
	.090	34

* *denotes alpha less than .05*

reliable information items are, as shown, much lower than the regression would predict.

It is possible that these two points represent only a change in the slope of the relationship between complete and tailored reliability, but it seems unlikely that such a sharp drop in tailored reliability would occur between r = .68 and r = .78 complete test reliability without the additional burdens of the guessing probability of .25 and a very heterogeneous item pool. Until additional experience accumulates, such speculation cannot be confirmed.

Among the secondary analyses a significant negative correlation was observed between number of items taken and the examinee's z score. There are two characteristics of the data which make this result plausible. First many of the brighter examinees took fewer items simply because of the ceiling of difficulty. Secondly, the items of lesser difficulty appear to be closer together on the continuum of ability-difficulty and therefore the program found more items at the ability level of the duller subjects.

The average values of c_{t3} found here for both tailored and complete data are lower than those found for the first live testing data. The averages for the first study were .16 and .42 for complete and tailored data. For the current data they are, respectively .12 and .22 .

Although TAILOR:APL failed to make successful use of the greater consistency in the tailored data, the existence of an index of tailored consistency may help to improve future tests. In spite of the fact that woefully unreliable scores were obtained in this study the program presented only an average of .41 of the items to each person. If the stringency of the statistical rules for establishing dominance relations were made to vary with c_{t3} , then perhaps when the data reached this low level of consistency, more relations could be required from each subject.

There are two characteristics of these items which may make them rather inappropriate for tailoring. First, they are moderately heterogeneous in subject matter, perhaps particularly in a college population where individuals have had varying amounts of formal course work in these subject areas. Thus, they may form subscales, and implications based on relative difficulty are subject to error.

Second, these are multiple choice items. The implied orders system is likely to be particularly vulnerable to the errors introduced by correct guesses. Some confirmation of this is found in Figure 5. Note first the band of "transition" responses, mixtures both correct and incorrect, along the main diagonal. There

is a considerable scattering of "boxes" (representing correct response) in the lower left portion, fairly far from this diagonal, representing very inconsistent correct responses. In contrast, there are only a few stars in the upper right, representing very inconsistent errors. Thus, this item pool may represent one which is inappropriate for tailoring.

The fact that more consistent data is obtained in tailored tests than complete tests raises the interesting possibility that test items are becoming more discriminating through some feature of tailoring. This means not only that items are not independent units with parameters that characterize them regardless of their surroundings, but also that item orders can be optimized to increase the efficiency of testing beyond the more generally expected effects of tailoring.

If increases in consistency continue to be observed when tests are tailored, explanation of this and optimization to make use of it will need to be investigated, both from a practical viewpoint and to gain a better understanding of the psychological process of problem-solving.

Group Testing Approaches

Overview

The group testing method of implied orders testing has been used in a variety of Monte Carlo and simulation studies. During the past year we have also tested several hundred introductory psychology students on three different batteries of tests, and have developed two separate versions, one in FORTRAN for the IBM 370, and another in APL at the Berkeley Management Science Center (MSC). In this section the current thinking regarding group testing approaches is presented. In general we conclude that, although the original presentation of the model has been in the group testing style, this approach has several serious practical limitations. Then a description of a second generation strategy is outlined which incorporates the best of the individual and group methods into one approach which can be used in a variety of settings.

The Group Testing Algorithm

The transition from a Monte Carlo program to a program for actual on-line

testing occurred in a straightforward manner. During the extensive series of Monte Carlos it was assumed that each simulated subject would receive one item each round until the completion of the test. If two or more batches of items were assessed in a single simulation, then all information from previous tests had to be completed before subsequent tests began.

The experience at the MSC was moderately encouraging with respect to this system except that the mini-computers which form the basis of this installation were much too slow to provide a reasonable test of the process. However, it did appear to have the serious drawback that it did not allow subjects to progress at different rates.

Therefore, the program was revised to follow the diagram given in Figure 6. If five subjects took a test for example, six operating programs were required, five of which were for the individual examinees, while the sixth was for the supervisor routine. These routines were designated INDTEST and MONITOR. INDTEST performed simple tasks such as displaying an item at a subject's terminal, scoring the response and communicating with MONITOR. MONITOR performed all the computations of the implied orders model and in addition policed the several individual sessions. The most crucial aspect of this system was the approach for communication between MONITOR and each INDTEST. From the perspective of INDTEST this meant that a subject could only answer an item when MONITOR supplied one; on the other hand, it was only efficient for MONITOR to proceed when approximately half the subjects had responded.

In practice this approach still translated into a large amount of wait time for all concerned, and although college students are fairly patient people, we have become extremely dissatisfied with it. Furthermore, the entire system is interlocked, such that if one computer terminal fails or one subject accidently presses a wrong button, the whole session halts. The amount of wait time and the number of restarts after a terminal failure have indicated two things about this system. First, from a practical point of view, the current implementation for group testing is basically suboptimal. It uses an unrealistic assumption about how subjects behave, for no group of individuals ever respond at exactly the same rates. It is also inefficient in terms of computer terminals because one machine must be reserved as a port for the MONITOR. Because there has been so much procedural manipulation and subject wait time, we feel that the group testing results which we have obtained on this system with human subjects are, unfortunately, misleading, except for some general procedural conclusions.

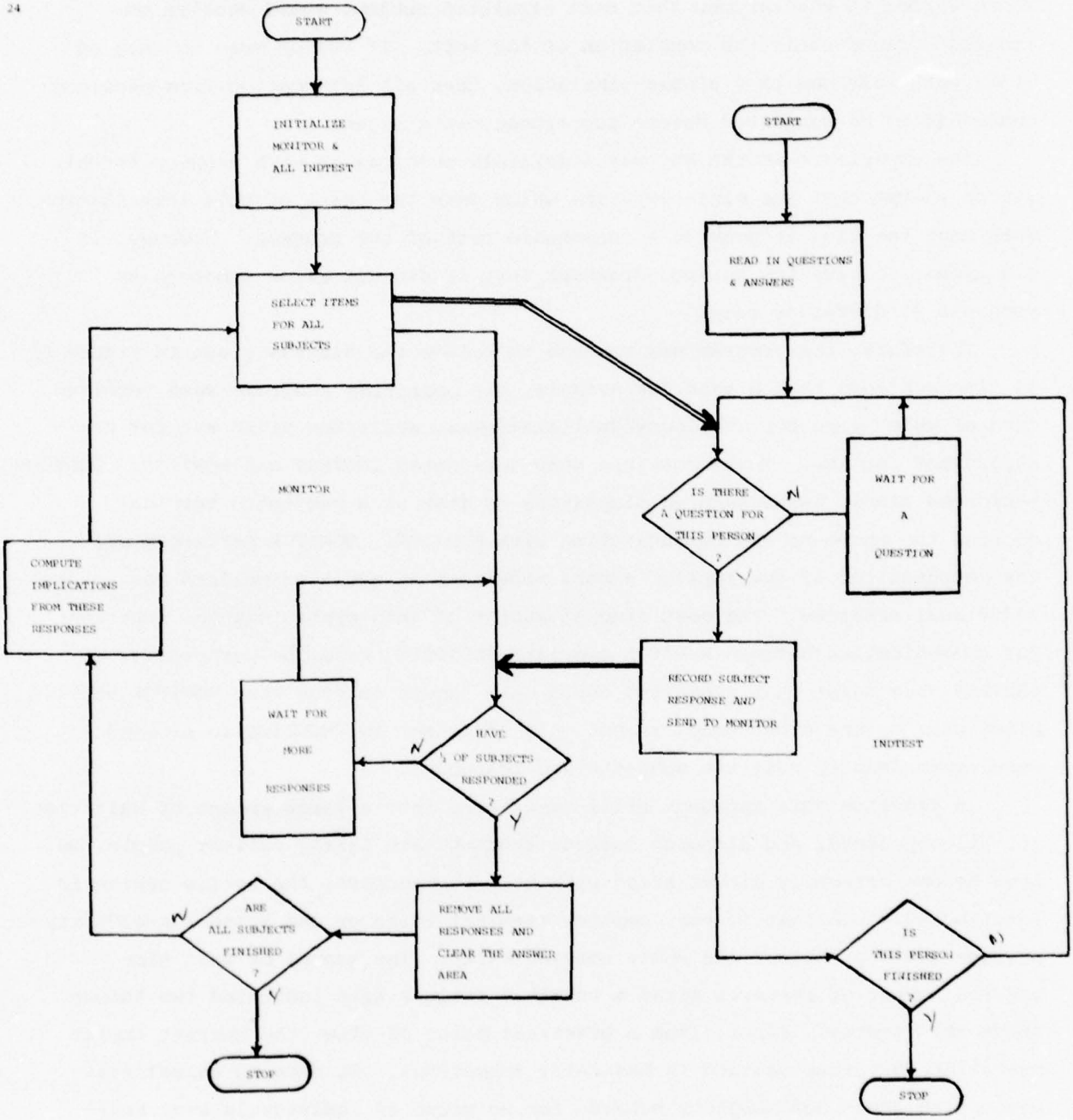


FIGURE 6: First generation TAILOR for group testing. Heavy lines indicate communication between MONITOR and INDTEST.

Although the above difficulties indicate that the recent research with the group approach has severe limitations, we cannot resist discussion of a consistently positive finding during the last year. In previous reports of the individual testing approach it was noted that a few complete tests must be given before a reduction in the number of items occurs, but that later subjects take tests of increasingly shorter lengths. The same pattern has been observed with the current group strategy, however, the saving in items begins with the very first group. Figure 7 indicates the extent to which this is possible. The plot labeled Group 1 shows the percentage of implications as a function of the percentage of responses for the first 5 subjects, using a well known test battery. The Group 12 plot shows how the number of responses decreases as a function of the number of previously tested subjects. In this instance, 46 subjects were tested prior to the final group of 5 subjects in Group 12. As can be seen, the savings is dramatic. The final group used only 37% of the items on the average. This figure compares favorably with previous findings from the individual testing data, and in fact is similar to the Monte Carlo studies for approximately the same numbers of persons and items (see Cudeck, et al, Note 5), We have benefited greatly from the information these practical considerations have provided, although the loss of statistical evaluation for the method is disappointing.

These concerns have resulted in a hybrid concept of the way to carry out implied orders testing. Although no working program has been written yet, such a routine will probably be developed in the future. It is basically a modification of the individual testing strategy, with some aspects of the supervision of the group approach. It is pictured schematically in Figure 8. Each individual program just makes a copy of the current contents of the integer item dominance matrix, while the supervisor prohibits one program from reading the data when another is simultaneously writing it back. The testing proceeds in a manner described by McCormick (Note 6) until the end of the test. At that time, the supervisor again controls the use of the common integer item dominance matrix so that individual programs record their data one at a time. This simple modification represents the best of each method used so far. The result is that subjects can work at their own pace, the number of persons being tested is, as the group procedure, limited only by the number of available terminals, and there is no need for a separate MONITOR session. Furthermore, it appears possible to write such an algorithm in either APL or FORTRAN on the current USC computer system. The modification to either program would be straightforward.

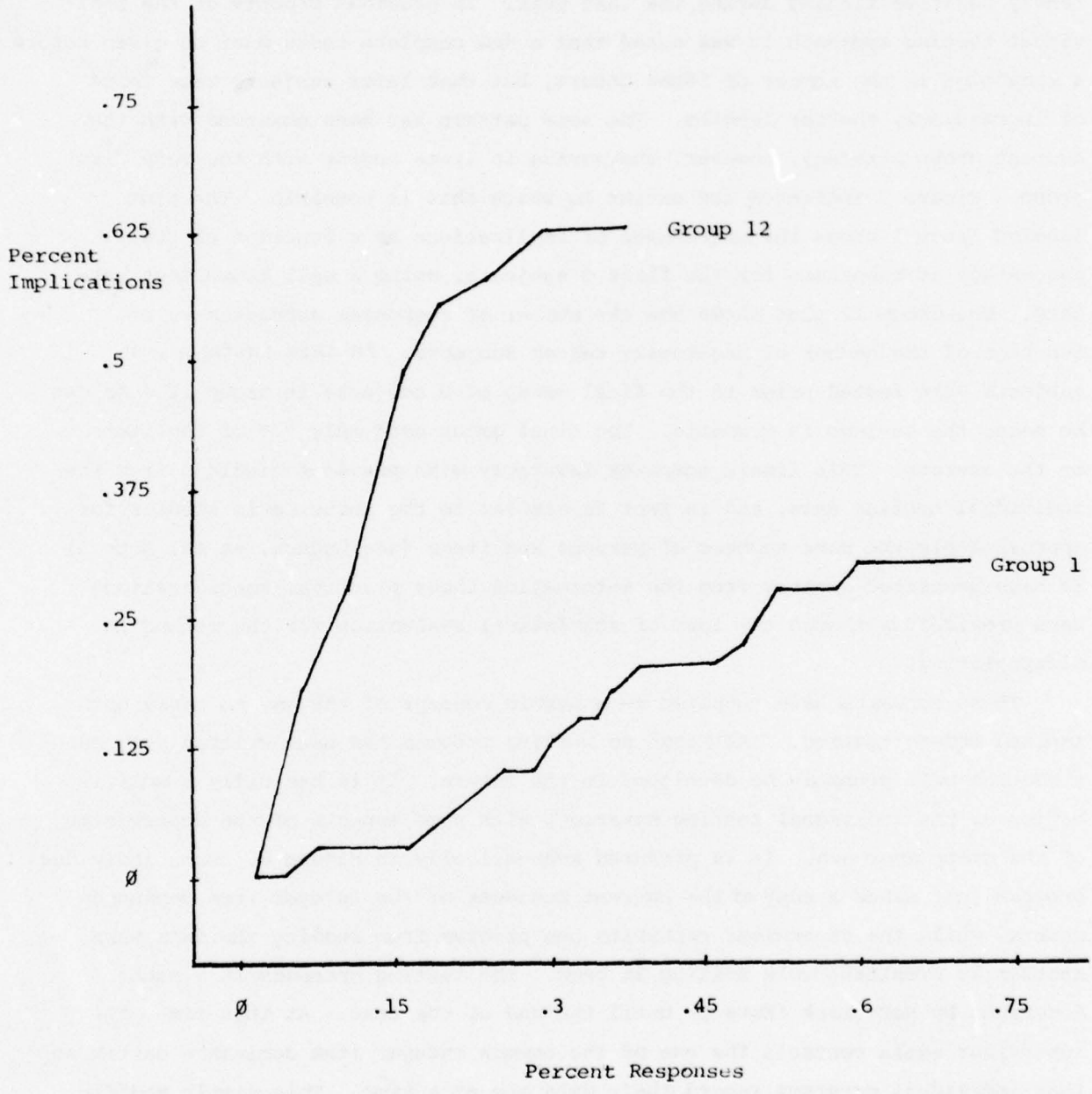


FIGURE 7: A comparison of percentage of implications as a function of responses for the first and twelfth groups.

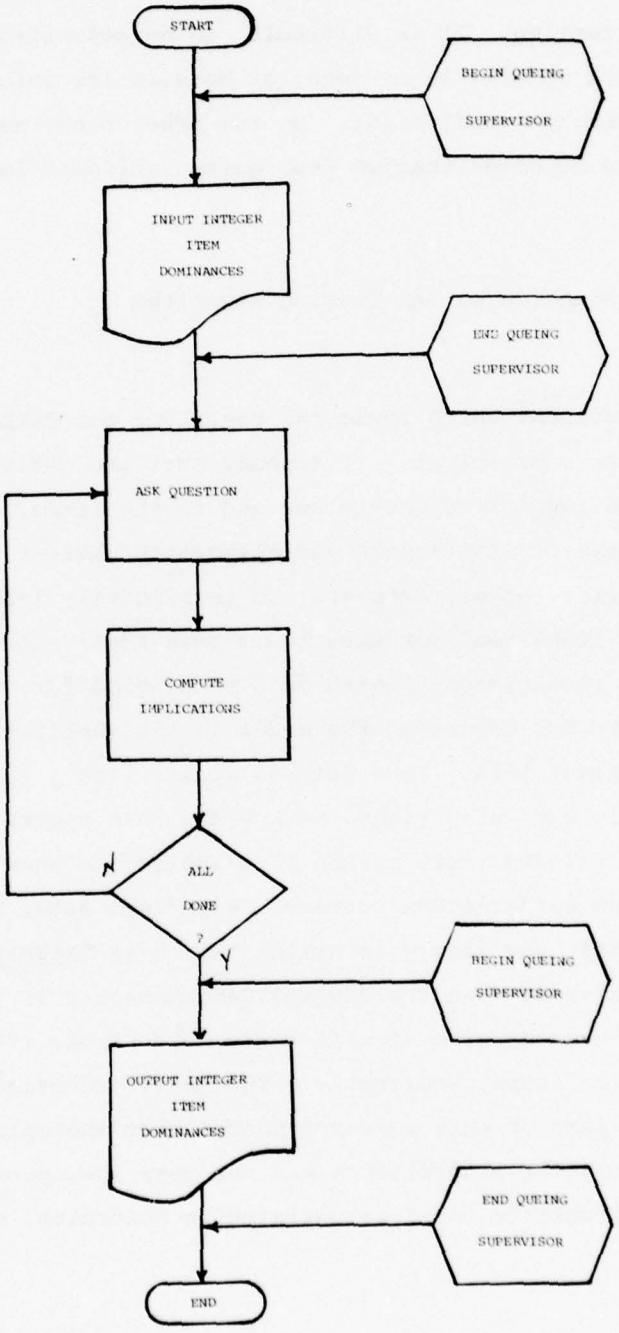


FIGURE 8: Second generation TAILOR for group testing.

IV. CONCLUSION: PROGRESS WITH IMPLIED ORDERS TAILORED TESTING

In this section, we present a final summary of the current outlook for Implied Orders tailored testing. It is difficult to be definitive in some areas because so much work remains to be done, or because the information which is available is somewhat conflicting. On the other hand, several findings have been replicated so often that we feel quite confident in making a general statement.

Evaluation of the Testing Algorithm

The Model

The implied orders concept which forms the basis for the TAILOR procedure takes the Guttman Scale as a prototype. It assumes that the goals of a test is to order the persons with respect to each other and to the items, as sketched in the introductory remarks of this report and elaborated earlier (Cliff, 1975; Cudeck, et al, Note 5; Cliff, et al, Note 4). No meaningfully large collection of data from persons and items conforms exactly to this ideal, however, and the data always contains inconsistency which must be allowed for.

The primary procedure for adjusting the model to the realities of the data is an approximate significant test. In a Guttman scale, item j is easier than k if there are more people who get j right and k wrong than the reverse. Similarly, in an error-free tailored test, person i is implied to answer item j correctly if he has gotten harder items correct. With real data, we substitute statistical comparisons that say item j is easier than k is "significantly more" persons get j right and k wrong than the reverse, and person i is implied to get item j correct if he has answered significantly more harder items correctly than he has answered easier items incorrectly. The early tinkering which as a very time-consuming part of this research showed that the optimal results occurred when the criterion for significance was set very low, provided certain safeguards (the "probability test" illustrated in McCormick, Note 6) were included.

Data on Implied Orders

The conclusions that seems justified from the experience to date is that such a model, bolstered by a heuristic device to absorb error, i.e., the significance test, works very well provided the data are highly consistent. The basis for this conclusion is in the Monte Carlo studies and the anagram data from TAILOR-APL.

The extensive Monte Carlo study, supported by the simulation with the file of Binet data (Cliff, et al, Note 4) showed that TAILOR scores would be more reliable than complete test scores based on the same number of items. This superiority would be appreciable if the items had very high discrimination indices, and negligible if they were moderate. This finding was bolstered by the startling results of the first anagram study (McCormick, Note 6). This found, in two separate replications, that TAILOR scores based on less than half the items were more reliable than total test scores, although not significantly so. Examination of the findings suggested that these data matrices were highly consistent, particularly in the case of the tailored data. The apparent fact that the tailored responses were more consistent than conventional responses to the same items was interesting in its own right and led to optimism that the TAILOR system could be readily made the basis for an operating system.

The second subsequent, real-subject trial reported here was not encouraging. It reported lower reliability for the TAILOR scores, and such results would in general be regarded as unsatisfactory for operational use. Reasons for this finding have been presented in the description of the study. The basic one is perhaps simply that TAILOR is sensitive to the degree of consistency of the data, and this test was simply not internally consistent enough for the TAILOR process to be effective. Possible multifactor structure and the presence of guessing are likely causes of the inconsistency.

Thus, the two tryouts with real data and Binet simulation (Cliff, et al, Note 2) show that our procedure can work as a tailoring system.

Computer Programs

The original idea for TAILOR was that the testing process would involve a group of subjects being tested at the same rate, in rounds, as it were. This is clearly inefficient since some subjects work faster than others and some will require more items than others. It is preferable that they be able to work at their own rate, or at least at one of several different rates. This is, of course, the way the individual program operates. However, in that program only one subject can be tested at a time; information from examinees is part of the data base only after they have completed the test. This too is undesirable.

An optimum version of the program is being contemplated which will in effect store the data from all subjects centrally, update it as it comes in from subjects and will handle several of them at the same time.

The centrally important data in the implied orders system is the matrix

which we call the item dominance matrix. It is an item by item matrix, denoted N in the theoretical articles (Cliff, 1975) and it has two forms, integer and binary (dichotomous). The integer form records in cell j,k the number of persons who get j wrong and k right. The binary form contains a 1 in the corresponding element if "significantly" more persons got j wrong and k right than the reverse. The significance tests converts the integer information into the binary form in the earlier versions of the program. The proposed version stores both.

The current version of the program calculates the integer dominance matrix as a matrix multiplication, after each round in the group version and after each person in the individual one. This is a very substantial calculation, even by the method we use. However, the response of a person to an item can only alter elements in one row or column of the matrix. This calculation is a singly subscripted loop -- at worst -- rather than a triply subscripted one, and so it can take place with great rapidity, enabling the program to handle inputs from a number of terminals. The program would operate on an as needed basis , so examinees could be taking the test at either the same time -- up to the limits of the terminal-monitoring capability -- or at different times. The program will tailor on the basis of the information that it has at that moment.

One final note may be made regarding data consistency and program operation. As currently written, TAILOR makes three kinds of implications and the decision rule in each instance is the same. One possible modification may be to alter the rules used according to the task,and in addition, to alter each rule according to the quality of the data. Thus if the items in a particular test are consistent, the requirements for an implication to be made could be relaxed. If the data are inconsistent, the decision rules could be made more stringent. As regards the kind of implication being made, it appears likely that item-item dominances which are based on many responses suggest a lenient decision criterion because the weight of evidence should serve to make any implication probably correct. Item-person implications may require a stringent rule due to the fact that much less information exists for the subjects and therefore any implications are more probabilistic.

Thus two different kinds of significance may be required, and they might be differently affected by the level of consistency of the data. Neither change would complicate the operation of the program very much, and they would hardly affect running speed. What they do require is research, particularly theoretical research, to establish the relation between the degree of consistency and optimum degree of tailoring.

Contributions to Test Theory

Consistency Measures

This research program also produced significant conceptual developments. The consistency measures proposed by Cliff (1975, 1977) provided a method of reconceptualizing test item consistency in terms of dominance concepts that are perhaps more valid to testing than are the traditional correlational ones. More directly important for tailored testing was the fact that these consistency measures were structured in such a way as to make them applicable to tailored tests.

Only recently has it been possible to incorporate these indices into the TAILOR program in order to monitor the consistency of the data. The final study reported above makes use of this information, but what is still developing is any intuitive feel for the magnitude of the numbers and what a given consistency value is likely to mean in terms of the efficiency of the process of tailoring. We anticipate that other workers will begin to reference these indices as means of judging consistency. One difficulty may be that while they are very plausibly defined they appear as startlingly low numbers to one used to the usual reliability coefficients.

Defining Subpools

The use of these coefficients to subdivide an item pool into homogeneous subsets was pioneered by Reynolds (Note 7). This proved very effective in his case, and these methods would probably be fruitful to pursue as a means of "factoring" dichotomous items. This is a direction that we intend to followup.

This development should be facilitated by an extension of the theoretical concepts which have been made only recently. Our work draws heavily on the concepts of dominance: item-person dominance, item-item dominance, and person-person dominance. The recent development is in essence a way of joining the concept of dominance to the traditional one of correlation. Relations on items may be redundant to relations on other items, contradictory to relations on other items, or unique to these items. Statistical definitions of these concepts allows one to consider dominance and correlation simultaneously. This fact suggests a means of measuring not just consistency but effectiveness of a pool of items, and also a method for dividing items into sub-pools of maximum effectivenss. Many details remain to be worked out, but the direction seems very promising and some of it has been presented at the Psychometrics Meeting in Uppsala. The measures suggested by these methods, too, generalize quite

straightforwardly to the tailored case.

Research Methodology in Tailored Testing

If one is developing a tailored testing system, one would of course like to know how well it works. One of the contributions of this project is felt to be the clear definitions used of tailored test effectiveness. These are not unusually subtle but they are somewhat different than used elsewhere. They use elementary design principles of not confounding the dependent with the independent variables, and not confounding independent variables with each other.

The two traditional qualities used to judge a psychometric variable are reliability and validity. Thus we have chosen to use as our major dependent variables the correlation of a tailored score with a parallel score or with a true score. Moreover, we have avoided the experimental confounding of tailored scores with either one. Thus, we derived in the real data studies two parallel tailored scores by independently tailoring two subtests and correlating the scores on the two. This practice avoids the confounding introduced by such practices as correlating a tailored score with a conventional score on the same items, where responses to items enter into the determination of both scores. Also, in our Monte Carlo studies where a true-score was used, the latent parameters are used to generate item scores only. The tailored test scores can then be correlated with the true scores to determine validity, and this validity is not confounded by any experimental dependence between the two.

The use of scores on randomly parallel tailored tests and randomly parallel complete tests to determine reliability also avoids the possible confounding of item quality with tailoring. This can occur when "tailoring" includes winnowing an item pool to select the best items. Here, measurement of efficiency of the tailoring process is confounded with the effects of using the more effective items. An appropriate question in evaluating a tailored process is "How valid is a score derived from tailoring this item pool compared to using the complete pool?"

The model used here automatically circumvents an inferential problem which troubles those tailoring procedures which estimate a true score and use an information measure as the definition of the effectivenss of the testing process. Aside from the fact that the item statistics used in this are at best estimates, not the true values of the item parameters, it appears likely that the items may behave differently in the tailored and untailored contexts, therefore, item

statistics derived from the one are not appropriate for exact estimation in the other. In any event, where such procedures are used, the criterion of correlation with separately and independently tailored tests should be used with real data evaluations. In Monte Carlo evaluations with an estimation process, one should be sure to use sample estimates of item parameters, not the population values, and actually correlate the estimated scores with true scores, not rely on information measures.

Future Prospects

Tailored Testing

It is perhaps surprizing that the implied orders system has worked at all. How well it works compared to other tailoring systems is hard to say since they have rarely provided the kind of data we feel is necessary for evaluation. It may be that no tailoring system will always work well except perhaps in a context which justifies a large investment in item development and pretesting. However, it seems that it should also work in situations where goals are such things as placement in a training sequence (i.e., the ability scale is large) or multifactor testing where one is primarily interested in identifying extremes on these factors. In both of these situations the Tailor program or a descendant of it would have an important place.

The suggested form of an operational program has been sketched earlier. It also seems that it might be desirable to incorporate response-timing features into the program so that in effect each item acts as several. Also, it might be useful where multiple choice responses are used for each alternative to be scored separately.

Ordinal Test Theory

The concepts used in this research have provided a basis for the measurement of consistency of items which is measured directly from the inter-relationships among the items. Thus tailored data can be evaluated as well as conventional. As indicated above, it appears that several new directions can be explored from these bases, and they are likely to be particularly useful for tailored testing. They also suggest new bases for determining the multifactor structure of data.

REFERENCE NOTES

1. Bock, R. D. Discussion participant in Weiss, D (ed.) Computerized adaptive trait measurement: Problems and prospects. Research Report 75-5, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1975.

2. Cliff, N. A basic test theory generalizable to tailored testing (Technical Report 1). Los Angeles, California: University of Southern California, Dept of Psychology, 1975.

3. Cliff, N., Cudeck, R. and McCormick, D. An empirical evaluation of implied orders as a basis for tailored testing. Paper presented at the Second Conference on Computerized Adaptive Testing, Minneapolis, Minnesota, 1977, in press.

4. Cliff, N., Cudeck, R. and McCormick, D. Evaluations of implied orders as a basis for tailored testing using simulations. (Technical Report 4) Los Angeles, California: University of Southern California, Dept of Psychology, 1977.

5. Cudeck, R., Cliff, N., Reynolds, T. and McCormick, D. Monte Carlo results from a computer program for tailored testing. (Technical Report 2) Los Angeles, California: University of Southern California, Dept of Psychology, 1976.

6. McCormick, D. TAILOR-APL: An interactive computer program for individual tailored testing. (Technical Report 5). Los Angeles, California: University of Southern California, Dept of Psychology, 1978.

7. Reynolds, T. The analysis of dominance matrices: Extraction of unidimensional orders within a multidimensional context. (Technical Report 3) Los Angeles, California: University of Southern California, Dept of Psychology, 1976.

8. The College Qualification Tests. Test 1. New York: The Psychological Corporation, 1956.

REFERENCES

Birnbaum, A. Some latent trait models and thier use in inferring an examinee's ability. In Lord, F. and Novick, M. Statistical theories of mental test scores. (Part 5) Reading, Mass: Addison-Wesley, 1968.

Cliff, N. Complete orders from incomplete data: Interactive ordering and tailored testing. Psychological Bulletin, 1975, 82, 289-302.

Cliff, N. A theory of consistency of ordering generalizable to tailored testing. Psychometrika, 1977, 42, 375-399.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cudeck, R., Cliff, N. and Kehoe, J. TAILOR: A FORTRAN procedure for interactive tailored testing. Educational and Psychological Measurement, 1977, 37, 767-769.

Knuth, D. F. The art of computer programming: Sorting and searching (Vol. 3) Reading Mass.: Addison-Wesley, 1973.

Loevinger, J. A systematic appraoch to the construction and evaluation of tests of ability. Psychological Monographs, 1947, 61 (4, Whole No. 285)

McCormick, D. and Cliff, N. TAILOR-APL: An interactive computer program for individual tailored testing. Educational and Psychological Measurement, 1977, 37, 771-774.

APPENDIX

Method for computing c_{t2} and c_{t3}

In the original presentation of these measures, c_{t2} is derived in a manner similar to many other psychometric devices in which the statistic is a ratio between explained variance to total variance. That is, c_{t2} uses two kinds of information from the ordered integer item dominance matrix, N = S'S, where S is as usual the n persons by x items score matrix. Then V, the total number of dominances, is

$$V = \sum_{j=1}^{x} \sum_{j=1}^{x} n_{jk} .$$

and v_m , the number of persons who respond in accordance with the order is

$$V_m = \sum_{j=1}^{x-1} \sum_{k=j+1}^{x} (n_{jk} - n_{kj}) .$$

The upper triangular portion of N will contain all the nonzero entries when S displays a perfect order. Thus V_m/V will be unity for data which conform to a Guttman scale, but zero for a random response pattern. Actually c_{t2} is modified to take on values in the range $-1 \leq c_{t2} \leq 1$ by the linear transformation.

$$c_{t2} = 2 (V_m/V) - 1 .$$

As can be seen, c_{t2} is similar in intent to a percentage agreement between several judges. It is equivalent to an index from Loevinger (1947) which was devised to assess homogeneous tests, a concept very closely related to the present development.

In contrast, c_{t3} relies on the consideration of marginal scores to determine consistency. When items are highly discriminating and have a broad range of difficulties, their marginal scores will reflect this condition, and one may expect highly consistent data for this reason alone. Thus the intention of c_{t3} is to correct for spuriously high consistency due to differences in difficulty. In that manner it resembles Cohen's kappa (Cohen, 1960) by correcting in accordance with the marginals. This adjustment is

$$c_{t3} = \frac{V_c - V}{V_c - V_m}$$

where

$$V_c = \sum_{j}^{x} \sum_{k \neq j}^{x} \hat{n}_{jk}$$

and

$$\hat{n}_{jk} = E(n_{jk}) = \frac{w_j (n - w_k)}{n}$$

with $w_j = \sum_{i=1}^{n} s_{ij}$.

The following detailed example is from Cliff (1977, p. 377) with inconsistent complete data.

Step 1. From S, compute $\tilde{S}'$, N and w_j .

$$S = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \tilde{S}' = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \qquad N = \begin{bmatrix} 0 & 2 & 2 & 2 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{bmatrix}$$

$$w_j = \begin{matrix} 2 & 3 & 3 & 3 \end{matrix}$$

Step 2. Using N, compute V and V_m .

$$V = 0 + 2 + 2 + 2 + 1 + 0 \ldots + 2 + 1 + 0 = 17.$$

$$V_m = (2 - 1) + (2 - 1) + (2 - 1) + (1 - 1) + (2 - 2) + (1 - 1) = 3$$

Step 3. Compute c_{t2}

$$c_{t2} = 2 (V_m/V) - 1 = 2(3/17) - 1 = -.647 .$$

Step 4. From the marginal scores, compute the expectation for N, $\hat{N}$.

$$\hat{N} = \begin{bmatrix} 6/5 & 9/5 & 9/5 & 9/5 \\ 4/5 & 6/5 & 6/5 & 6/5 \\ 4/5 & 6/5 & 6/5 & 6/5 \\ 4/5 & 6/5 & 6/5 & 6/5 \end{bmatrix} \qquad \begin{matrix} (n - w_j) \\ 3 \\ 2 \\ 2 \\ 2 \end{matrix}$$

$$w_j = \begin{matrix} 2 & 3 & 3 & 3 \end{matrix}$$

Step 5. Calculate V_c

$$V_c = 9/5 + 9/5 + 9/5 + 4/5 + 6/5 + \ldots + 6/5 + 6/5 = 15$$

Step 6. Find c_{t3} in

$$c_{t3} = \frac{V_c - V}{V_c - V_m} = \frac{15 - 17}{15 - 3} = -.167 \quad .$$

Distribution List

Navy

4 Dr. Marshall J. Farr, Director
Personnel and Training Research Programs
Office of Naval Research (Code 458)
Arlington, VA 22217

1 ONR Branch Office
495 Summer Street
Boston, MA 02210
ATTN: Dr. James Lester

1 ONR Branch
1030 East Green Street
Pasadena, CA 91101
ATTN: Dr. Eugene Gloye

1 ONR Branch Office
536 South Clark Street
Chicago, IL 60605
ATTN: Dr. Charles E. Davis

1 Dr. M. A. Bertin
Scientific Director
Office of Naval Research
Scientific Liaison Group/Tokyo
American Embassy
APO San Francisco 96503

1 Office of Naval Research
Code 200
Arlington, VA 22217

6 Director
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Technical Director
Navy Personnel Research and Development Center
San Diego, CA 92152

1 Assistant for Research Liaison
Bureau of Naval Personnel (Pers Or)
Room 1416, Arlington Annex
Washington, DC 20370

1 Assistant Deputy Chief of Naval Personnel for Retension Analysis and Coordination (Pers 12)
Room 2403, Arlington Annex
Washington, DC 20370

1 CDR Paul D. Nelson, MSC, USN
Naval Medical R & D Command (Code 44)
National Naval Medical Center
Bethesda, MD 20014

1 Commanding Officer
Naval Health Research Center
San Diego, CA 92152
ATTN: Library

1 Chairman
Behavioral Science Department
Naval Command & Management Division
U. S. Naval Academy
Annapolis, MD 21402

1 Dr. Jack R. Borsting
U. S. Naval Postgraduate School
Department of Operations Research
Monterey, CA 93940

1 Director, Navy Occupational Task Analysis Program (NOTAP)
Navy Personnel Program Support Activity
Building 1304, Bolling AFB
Washington, DC 20336

1 Office of Civilian Manpower Management
Code 64
Washington, DC 20390
ATTN: Dr. Richard J. Niehaus

1 Superintendent
Naval Postgraduate School
Monterey, CA 93940
ATTN: Library (Code 2124)

1 Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054
ATTN: Dr. Norman J. Kerr

1 Principal Civilian Advisor
for Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508
ATTN: Dr. William L. Maloy

1 Director
Training Analysis & Evaluation Group
Code N-00t
Department of the Navy
Orlando, FL 32813
ATTN: Dr. Alfred F. Smode

1 Navy Personnel Research
and Development Center
Code 01
San Diego, CA 92152

5 Navy Personnel Research
and Development Center
Code 02
San Diego, CA 92152
ATTN: A.A. Sjoholm

2 Navy Personnel Research
and Development Center
Code 310
San Diego, CA 92152
ATTN: Dr. Martin F. Wiskoff

1 Navy Personnel Research
and Development Center
San Diego, CA 92152
ATTN: Library

1 Navy Personnel Research
and Development Center
Code 9041
San Diego, CA 92152
ATTN: Dr. J. D. Fletcher

1 D. M. Gragg, CAPT, MC, USN
Head, Educational Programs Development Department
Naval Health Sciences Education and
Training Command
Bethesda, MD 20014

Army

1 Technical Director
U. S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Headquarters
U. S. Army Administration Center
Personnel Administration Combat
Development Activity
ATCP- HRQ
Ft. Benjamin Harrison, IN 46249

1 Armed Forces Staff College
Norfolk, VA 23511
ATTN: Library

1 Dr. Stanley L. Cohen
U. S. Army Research Institute for
the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Ralph Dusek
U. S. Army Research Institute for
the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Joseph Ward
U. S. Army Research Institute for
the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 HQ USAREUR & 7th Army
ODCSOPS
USAREUR Director of GED
APO New York 09403

1 ARI Field Unit - Leavenworth
Post Office Box 3122
Fort Leavenworth, KS 66027

1 Dr. Ralph Canter
U. S. Army Research Institute for
the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Milton Maier
U. S. Army Research Institute
for the Behavioral and Social
Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Milton S. Katz, Chief
Individual Training & Perfor-
mance Evaluation
U. S. Army Research Institute for
the Behavioral and Social
Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Air Force

1 Research Branch
AF/DPMYAR
Randolph AFB, Tx 78148

1 Dr. G. A. Echstrand (AFHRL/AST)
Wright Patterson AFB
Ohio 45433

1 AFHRL/DOJN
Stop #63
Lackland AFB, Tx 78236

1 Dr. Martin Rockway (AFHRL/TT)
Lowry AFB
Colorado 80230

1 Dr. Alfred R. Fregly
AFOSR/NL
1400 Wilson Boulevard
Arlington, Va 22209

1 AFHRL/PED
Stop #63
Lackland AFB, Tx 78236

1 Major Wayne S. Sellman
Chief of Personnel Testing
HQ USAF/DPMYP
Randolph AFB, Tx 78148

Marine Corps

1 Director, Office of Manpower Utilization
Headquarters, Marine Corps (Code MPU)
MCB (Building 2009)
Quantico, VA 22134

1 Dr. A. L. Slafkosky
Scientific Advisor (Code RD-1)
Headquarters, U.S. Marine Corps
Washington, DC 20380

1 Chief, Academic Department
Education Center
Marine Corps Development and
Education Command
Marine Corps Base
Quantico, VA 22134

1 Mr. E. A. Dover
2711 South Veitch Street
Arlington, VA 22206

Coast Guard

1 Mr. Joseph J. Cowan, Chief
Psychological Research Branch (G-P-1/62)
U.S. Coast Guard Headquarters
Washington, DC 20590

1 Mr. Thomas Warm
U.S. Coast Guard Institute
P.O. Substation 18
Oklahoma City, OK 73169

Other DOD

1 Dr. Harold F. O'Neil, Jr.
Advanced Research Projects Agency
Cybernetics Technology, Rm. 625
1400 Wilson Boulevard
Arlington, VA 22209

12 Defense Documentation Center
Cameron Station, Building 5
Alexandria, VA 22314
ATTN: TC

Other Government

1 Dr. Lorraine D. Eyde
Personnel Research and Development Center
U. S. Civil Service Commission
1900 E Street, N. W.
Washington, DC 20415

1 Dr. William Gorham, Director
Personnel Research and Development Center
U. S. Civil Service Commission
1900 E Street, N. W.
Washington, DC 20415

1 Dr. Vern Urry
Personnel Research and Development Center
U. S. Civil Service Commission
1900 E Street, N. W.
Washington, DC 20415

1 Dr. Harold T. Yahr
Personnel Research and Development Center
U. S. Civil Service Commission
1900 E Street, N. W.
Washington, DC 20415

1 Dr. Andrew R. Molnar
Technical Innovations in Education Group
National Science Foundation
1800 G Street, N. W.
Washington, DC 20550

1 U. S. Civil Service Commission
Federal Office Building
Chicago Regional Staff Division
Regional Psychologist
230 South Dearborn Street
Chicago, IL 60604
ATTN; C. S. Winiewicz

1 Dr. Carl Frederiksen
Learning Division, Basic Skills Group
National Institute of Education
1200 19th Street, N. W.
Washington, DC 20208

Miscellaneous

1 Dr. Scarvia B. Anderson
Educational Testing Service
17 Executive Park Drive, N. E.
Atlanta, GA 30329

1 Mr. Samuel Ball
Educational Testing Service
Princeton, NJ 08540

1 Dr. Gerald V. Barrett
University of Akron
Department of Psychology
Akron, OH 44325

1 Dr. Kenneth E. Clark
University of Rochester
College of Arts and Sciences
River Campus Station
Rochester, NY 14627

1 Dr. John J. Collins
Vice President
Essex Corporation
6305 Caminito Estrellado
San Diego, CA 92120

1 Dr. Rene V. Dawis
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 Dr. Marvin D. Dunnette
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 ERIC
Processing and Reference Facility
4833 Rugby Avenue
Bethesda, MD 20014

1 Major I. N. Evonic
Canadian Forces Personnel Applied Research Unit
1107 Avenue Road
Toronto, Ontario, CANADA

1 Dr. Victor Fields
Montgomery College
Department of Psychology
Rockville, MD 20850

1 Dr. Edwin A. Fleishman
Visiting Professor
University of California
Graduate School of Administration
Irvine, CA 92664

1 Dr. John R. Frederiksen
Bolt, Beranek and Newman, Inc.
50 Moulton Street
Cambridge, MA 02138

1 Dr. Robert Glaser, Co-Director
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15213

1 Dr. Richard S. Hatch
Decision Systems Associates, Inc.
5640 Nicholson Lane
Rockville, MD 20852

1 Dr. M. D. Havron
Human Sciences Research, Inc.
7710 Old Spring House Road
West Gate Industrial Park
McLean, VA 22101

1 HumRRO Central Division
400 Plaza Building
Pace Boulevard at Fairfield Drive
Pensacola, FL 32505

1 HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921
ATTN: Library

1 Dr. David Klahr
Carnegie-Mellon University
Department of Psychology
Pittsburgh, PA 15213

1 Dr. Alma E. Lantz
University of Denver
Denver Research Institute
Industrial Economics Division
Denver, CO 80210

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. Robert R. Mackie
Human Factors Research, Inc.
6780 Corton Drive
Santa Barbara Research Park
Goleta, CA 93017

1 Dr. William C. Mann
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina Del Rey, CA 90291

1 Mr. Edmond Marks
315 Old Main
Pennsylvania State University
University Park, PA 16802

1 Richard T. Mowday
College of Business Administration
University of Nebraska, Lincoln
Lincoln, NE 68588

1 Dr. Leo Munday, Vice-President
American College Testing Program
P. O. Box 168
Iowa City, IA 52240

1 Mr. Luigi Petrullo
2431 North Edgewood Street
Arlington, VA 22217

1 Dr. Steven M. Pine
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 Dr. Diane M. Ramsey-Klee
R-K Research & System Design
3947 Ridgemont Drive
Malibu, CA 90265

1 Dr. Joseph W. Rigney
University of Southern California
Behavioral Technology Laboratories
3717 South Grand
Los Angeles, CA 90007

1 Dr. Andrew M. Rose
American Institutes for Research
3301 New Mexico Avenue, N. W.
Washington, DC 20016

1 Dr. George E. Rowland
Rowland and Company, Inc.
P. O. Box 61
Haddonfield, NJ 08033

1 Dr. Benjamin Schneider
University of Psychology
Department of Psychology
College Park, MD 20742

1 Dr. Lyle Schoenfeldt
Department of Psychology
University of Georgia
Athens, Georgia 30602

1 Dr. Arthur I. Siegel
Applied Psychological Services
404 East Lancaster Avenue
Wayne, PA 19087

1 Dr. Henry P. Sims, Jr.
Room 630 - Business
Indiana University
Bloomington, IN 47401

1 Dr. C. Harold Stone
1428 Virginia Avenue
Glendale, CA 91202

1 Dr. Patrick Suppes, Director
Institute for Mathematical Studies
in the Social Sciences
Stanford University
Stanford, CA 94305

1 Dr. Sigmund Tobias
PH.D Programs in Education
Graduate Center
City University of New York
33 West 42nd Street
New York, NY 10036

1 Dr. David J. Weiss
University of Minnesota
Department of Psychology
N660 Elliott Hall
Minneapolis, MN 55455

1 Dr. K. Wescourt
Stanford University
Institute for Mathematical Studies
in the Social Sciences
Stanford, CA 94305

1 Dr. Anita West
Denver Research Institute
University of Denver
Denver, CO 80210

1 Mr. George Wheaton
American Institutes for Research
3301 New Mexico Avenue, N.W.
Washington, DC 20016